MULTIVARIATE NORMALITY GOODNESS OF FIT TESTS

By

Harold L. Crutcher
National Oceanic and Atmospheric Administration

And

Lee W. Falls
George C. Marshall Space Flight Center

MULTIVARIATE NORMALITY GOODNESS OF FIT TESTS

This program provides an extension of Pearson's chi-square (χ^2) goodness of fit test for univariate distributions to the multivariate normal models. Two dimensions (bivariate case) through five dimensions are included in the program. Extensions to higher dimensions may be made if required. The basis of the multivariate test is the fact that the exponential of the multivariate normal distribution is distributed as a chi-square variable with degrees of freedom equal to the numbers of dimensions.

The probability density function of the multivariate normal distribution may be written as

$$p(x_1, x_2, \ldots, x_\nu) = A e^{-\frac{1}{2}(Q)} \qquad (a)$$

for ν correlated random variates $(x_1, x_2, \ldots, x_\nu)$. A is a constant for a distribution and is a function of dimensions ν and the correlation matrix. The quadratic Q of equation (a) is distributed as chi-square with ν degrees of freedom. Equation (7) gives the formula for Q (in the program notation χ^2_o is Q).

The program also provides the Kolmogorov-Smirnov (KS) goodness of fit test. This is accomplished by computing the maximum absolute difference (MAD) between the theoretical multivariate normal distribution and the empirical distribution.

The user of this program should note that these multivariate tests are strictly valid only for independent samples and for multivariate distributions whose marginal distributions are unimodal. If these conditions are not met by the sample, the effects upon the tests are unknown. However, the authors believe that the tests may still be used subjectively even if the described restrictions are not met.

LIST OF SYMBOLS

Symbol	Definition
C	Correlation matrix
c	Constant
d.f.	Final degrees of freedom for test
k	Number of class intervals after grouping for the χ^2 test
k_N	Number of class intervals before grouping for the χ^2 test
ℓn	Natural logarithm
log	Common logarithm
N	Sample size
R	Correlation coefficient
S	Sample standard deviation
x	Variate
$\bar{x}$	Mean of x
D	Determinant of correlation matrix
χ^2	Chi-square
χ_o^2	Computed chi-square from quadratic
r_{ij}	Cofactor of R (elements of correlation matrix)
r^{ij}	Cofactor of R divided by determinant of correlation matrix
P	Probability
ν	Dimensions, number of variates
K	Number of tables, number of samples
MAD	Maximum absolute difference
$d.f._I$	Initial or desired degrees of freedom for test
F_o	Observed frequency
F_e	Expected theoretical frequency

LIST OF SYMBOLS (Concluded)

Symbol	Definition
x_B	Bottom of class interval
x_i	Class mark
x_T	Top of class interval
X^2	Calculated chi-square index
MINFE	Minimum expected frequency per class interval required for the chi-square test

FINAL INSTRUCTIONS TO THE COMPUTER

1. The following options will be included for any sample size N for this program:

Option 1: 2-dimensional case (2 variates or "pairs") designated as (x_1, x_2)

Option 2: 3-dimensional case (3 variates or "triplets") designated as (x_1, x_2, x_3)

Option 3: 4-dimensional case (4 variates) designated as (x_1, x_2, x_3, x_4)

Option 4: 5-dimensional case (5 variates) designated as $(x_1, x_2, x_3, x_4, x_5)$

We will describe the procedure to follow for samples of 11 pairs (option 1, 2 dimensions, N = 11). This same procedure will be performed for the other options when they are chosen.

2. For each of the samples of 11 pairs compute the following:

$$\bar{x}_1 = \frac{1}{N} \sum_{i=1}^{N} x_i \ , \qquad \bar{x}_2 = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{1}$$

$$s^2_{x_1} = \frac{N \sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i \right)^2}{N(N-1)} \ , \qquad s^2_{x_2} = \frac{N \sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i \right)^2}{N(N-1)} \tag{2}$$

$$s_{x_1} = \sqrt{s^2_{x_1}} \ , \qquad s_{x_2} = \sqrt{s^2_{x_2}} \tag{3}$$

$$R_{x_1, x_2} = \frac{N \sum_{i=1}^{N} (x_1)_i (x_2)_i - \sum_{i=1}^{N} (x_1)_i \sum_{i=1}^{N} (x_2)_i}{N(N-1) S_{x_1} S_{x_2}} \ . \tag{4}$$

Equation (4) gives the serial correlation coefficient R_{x_1, x_2} between the variables x_1 and x_2. The general equation for correlations will be

$$R_{x_\ell, x_j} = \frac{N \sum_{i=1}^{N} (x_\ell)_i (x_j)_i - \sum_{i=1}^{N} (x_\ell)_i \sum_{i=1}^{N} (x_j)_i}{N(N-1) S_{x_\ell} S_{x_j}} . \quad (5)$$

When $\ell = j$, then $R_{x_\ell x_\ell}$ or $R_{x_j x_j} = 1$; i.e., in the following correlation matrix, all diagonal elements = 1. (Note ℓ and j = 1, 2, 3, 4, 5 to include all options.) Using equation (5), form the correlation matrix C as follows:

Bivariate Case

$$C = \begin{bmatrix} R_{11} & R_{12} & R_{13} & R_{14} & R_{15} \\ R_{21} & R_{22} & R_{23} & R_{24} & R_{25} \\ R_{31} & R_{32} & R_{33} & R_{34} & R_{35} \\ R_{41} & R_{42} & R_{43} & R_{44} & R_{45} \\ R_{51} & R_{52} & R_{53} & R_{54} & R_{55} \end{bmatrix} \quad (6)$$

Note that this matrix is a symmetric matrix with $R_{12} = R_{21}$, $R_{13} = R_{31}$, etc. Consequently, our notation will be confined to the upper half of the correlation matrix, including the diagonal elements. Solve for the determinant of C; call this determinant D.

Now, the general equation for χ_o^2 is

$$\chi_o^2 = \frac{1}{D}\left[r_{11}\left(\frac{x_1 - \bar{x}_1}{S_{x_1}}\right)^2 + r_{22}\left(\frac{x_2 - \bar{x}_2}{S_{x_2}}\right)^2 + \cdots\right.$$

$$r_{55}\left(\frac{x_5 - \bar{x}_5}{S_{x_5}}\right)^2 + 2r_{12}\left(\frac{x_1 - \bar{x}_1}{S_{x_1}}\right)\left(\frac{x_2 - \bar{x}_2}{S_{x_2}}\right) +$$

$$2r_{13}\left(\frac{x_1 - \bar{x}_1}{S_{x_1}}\right)\left(\frac{x_3 - \bar{x}_3}{S_{x_3}}\right) + \cdots 2r_{15}\left(\frac{x_1 - \bar{x}_1}{S_{x_1}}\right)\left(\frac{x_5 - \bar{x}_5}{S_{x_5}}\right) \qquad (7)$$

$$+ 2r_{23}\left(\frac{x_2 - \bar{x}_2}{S_{x_2}}\right)\left(\frac{x_3 - \bar{x}_3}{S_{x_3}}\right) + \cdots 2r_{25}\left(\frac{x_2 - \bar{x}_2}{S_{x_2}}\right)\left(\frac{x_5 - \bar{x}_5}{S_{x_5}}\right)$$

$$\left. + 2r_{34}\left(\frac{x_3 - \bar{x}_3}{S_{x_3}}\right)\left(\frac{x_4 - \bar{x}_4}{S_{x_4}}\right) + \cdots 2r_{45}\left(\frac{x_4 - \bar{x}_4}{S_{x_4}}\right)\left(\frac{x_5 - \bar{x}_5}{S_{x_5}}\right)\right]$$

where r_{ij} is the cofactor of R_{ij} (the elements of correlation matrix C).

Let $r_{ij}/D = r^{ij}$. Now, the general equation for χ_o^2 becomes

$$\chi_o^2 = r^{11}\left(\frac{x_1 - \bar{x}_1}{S_{x_1}}\right)^2 + r^{22}\left(\frac{x_2 - \bar{x}_2}{S_{x_2}}\right)^2 + \cdots r^{55}\left(\frac{x_5 - \bar{x}_5}{S_{x_5}}\right)^2$$

$$+ 2r^{12}\left(\frac{x_1 - \bar{x}_1}{S_{x_1}}\right)\left(\frac{x_2 - \bar{x}_2}{S_{x_2}}\right) + 2r^{13}\left(\frac{x_1 - \bar{x}_1}{S_{x_1}}\right)\left(\frac{x_3 - \bar{x}_3}{S_{x_3}}\right)$$

$$+ \cdots 2r^{15}\left(\frac{x_1 - \bar{x}_1}{S_{x_1}}\right)\left(\frac{x_5 - \bar{x}_5}{S_{x_5}}\right) + 2r^{23}\left(\frac{x_2 - \bar{x}_2}{S_{x_2}}\right)\left(\frac{x_3 - \bar{x}_3}{S_{x_3}}\right) \qquad (8)$$

$$+ \cdots 2r^{25}\left(\frac{x_2 - \bar{x}_2}{S_{x_2}}\right)\left(\frac{x_5 - \bar{x}_5}{S_{x_5}}\right) + 2r^{34}\left(\frac{x_3 - \bar{x}_3}{S_{x_3}}\right)\left(\frac{x_4 - \bar{x}_4}{S_{x_4}}\right)$$

$$+ \cdots 2r^{45}\left(\frac{x_4 - \bar{x}_4}{S_{x_4}}\right)\left(\frac{x_5 - \bar{x}_5}{S_{x_5}}\right) \quad .$$

A standard matrix inversion program applied to correlation matrix C (the elements are R_{ij}) provides a matrix whose elements are r^{ij}. For this reason, equation (8) is used for the computation of χ_o^2.

3. (a) The option chosen in part 1 will determine the dimensions to be used in the general equations. As before, we will use option 1 (2 dimensions, samples of 11 pairs) for illustration.

(b) For N = 11, compute $\bar{x}_1$, $\bar{x}_2$, S_{x_1}, S_{x_2}, R_{11}, R_{12}, R_{21}, and R_{22} using equations (1), (2), (3), and (5). Note that $R_{x_1, x_2} = R_{12}$, $R_{x_1, x_3} = R_{13}$, $R_{x_1, x_4} = R_{14}$, $R_{x_2, x_3} = R_{23}$, $R_{x_2, x_4} = R_{24}$, etc., in matrix (6).

(c) Form the correlation matrix C. For the 2-dimensional case (bivariate) this will include elements R_{11}, R_{12}, R_{21}, and R_{22} only. Invert matrix C to obtain r^{11}, r^{12}, r^{21}, and r^{22}.

(d) Introducing the 11 observations along with the parameters computed in (3b) and (3c) into equation (8), we obtain 11 values of χ_o^2.

(e) Order the 11 values of χ_o^2 obtained in part (3d) according to increasing magnitude. Compute the empirical probability of exceeding χ_o^2 corresponding to the N = 11 ordered χ_o^2 values using

$$[1 - P(\chi_o^2)] = \frac{N - i - c + 1}{N - 2c + 1} \qquad (9)$$

for i = 1, 2, 3, . . . , 11.

$$c = -\frac{1}{4}\nu + \left(1 - \frac{4\,\ell n\, N}{N}\right) \; ; \qquad \begin{matrix} N > 2 \\ \nu = \text{dimensions} \end{matrix} .$$

(f) Prepare Table 1. This table will be the 11 ordered χ_o^2 values vs $[1 - P(\chi_o^2)]$. Note that there will be K table 1's corresponding to the K samples of size 11.

(g) Plot the ordered χ_o^2 values vs $[1 - P(\chi_o^2)]$ on linear vs $\log_{10}$ coordinates. Plot the theoretical fit to these points using the 1108 subroutine 12.2, CHI. Use the appropriate degrees of freedom (d.f.); i.e., for the 2-dimensional case, d.f. = 2, for the 3-dimensional case, d.f. = 3, etc.

(h) Goodness of Fit: Compute the theoretical cumulative probabilities $[1 - P(\chi^2)]$ for the N = 11 ordered χ_o^2 values using 1108 subroutine 12.2, CHI. Add these values of $[1 - P(\chi^2)]$ to Table 1. Compute the maximum absolute difference between $[1 - P(\chi_o^2)]$ and $[1 - P(\chi^2)]$. Call this quantity MAD. Note that there will be K values of MAD corresponding to the K samples of size N = 11.

(i) Using 1108 subroutine 13.4, CHIN, for the N = 11 values of $[1 - P(\chi_o^2)]$ in Table 1, calculate the corresponding values of χ^2 located on the theoretical straight line. Add these values of χ^2 to Table 1.

4. Goodness of Fit. For each sample of size N perform the following:

(a) Group the χ_o^2 values into equal class intervals. The number of class intervals is

$$k_N = d.f._I + \frac{(\nu+1)(\nu+2)}{2} \tag{10}$$

where k_N is the number of class intervals before grouping, $d.f._I$ is the desired degrees of freedom for the χ^2 test, and ν is the dimension of the sample. Tabulate the observed frequency (F_o) of the χ_o^2 values. Let

x_B = bottom of class interval

x_i = class mark

x_T = top of class interval

F_o = observed frequency

(b) Compute the theoretical cumulative probabilities $P(\chi^2)$ using 1108 subroutine 12.2 CHI. The input to the program will be x_T and appropriate dimensions (2, 3, 4, or 5).

(c) Compute the expected theoretical frequencies (F_e)

$$F_e = N[P(\chi^2)_i - P(\chi^2)_{i-1}] \quad .$$

(d) Compute X^2 using

$$X^2 = \sum_{i=1}^{k} \frac{(F_o - F_e)^2}{F_e} \quad .$$

If $F_e < 1.0$, adjacent F_e's must be added so that $F_e \geq 1.0$. If any F_e's are grouped to satisfy this rule, the corresponding F_o must be added in a similar manner. ($F_e < 1.0$) will be a variable input denoted by MINFE. Let k = number of class intervals after grouping for the χ^2 test. Now,

$$\text{d.f.} = \text{d.f.}_I - (k_N - k) \quad .$$

5. Printout Format:

Table 1		2 Dimensions		N = 11				
χ_o^2	$1-P(\chi^2)$	χ^2	$1-P(\chi_o^2)$	x_1	x_2	x_3	x_4	x_5
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.

$\bar{x}_1 =$, $\bar{x}_2 =$, ...

$S_{x_1} =$, $S_{x_2} =$, ...

$R_{11} =$, $R_{12} =$, $R_{21} =$, $R_{22} =$, ...

Table 2 2 Dimensions N = 11

x_i	x_T	F_o	$P(\chi^2)$	F_e
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

MAD = , X^2 = , k_N = , k = , d.f. = , MINFE = .

For further information regarding these multivariate normality tests and additional tests on the same subject, the reader is referred to NASA Technical Note TN D-8226 entitled "Multivariate Normality", May 1976 by Harold L. Crutcher and Lee W. Falls.